Smart Response System using Speech Emotion

Anuraag Advani, Ankita Gandhi, Harita Jagad, Prof. Dharmesh Mistry
Department of Computer Engineering
Dwarkadas.J.Sanghvi College of Engineering,Vile Parle(West), Mumbai
anuraag.advani@gmail.com,ankitagandhi.6694@gmail.com,haritajagad95@gmail.com,dharmeshkumar.mistry@djsce.ac.in

Abstract- Speech recognition is becoming pervasive in many areas like smart-assistants, navigation systems and home automation to name a few. To build a truly smart speech recognition system, not only the content but also the emotion embedded in the content should be extracted. Generating responses according to human's speech emotion is challenging as speech involves dealing with many ambiguous features. In this paper we propose to develop a system, which uses signal features like pitch, energy and Mel frequency cepstral coefficients, then uses Gaussian mixture models to detect human emotion which will further be used to generate output using extreme learning machine.

Keywords—*Gaussian Mixture Models; Extreme learning machine; emotion recognition; smart response generation*

*****

I. INTRODUCTION

Speech Emotion Recognition is a branch of Human Computer Interaction where the parts of speech called utterances are studied and the underlying emotion of the speaker is detected. We will use this technology, integrated with the Global Positioning System and other resources available in a vehicle to develop quick and interactive response generating system which is a smart in-car navigation system.

In this paper, we propose our smart in-car navigation system that emphasizes on human-machine interaction. We do so by implementing 'Speech Emotion Recognition', wherein our proposed system will have speech as input via a microphone and using this, the emotion of the speaker will be detected. It will be programmed to recommend different locations to navigate depending on the mood detected in the car. It can be used in emergencies to contact concerned authorities when it senses emotions like fear etc. It can also provide a form of entertainment by providing interaction or suggesting music. We also need to ensure that the functionality is non-intrusive and keeps the driver in control without hindering the driving experience. The proposed system will extract Mel Frequency Cepstral Coefficients, energy and pitch, and cluster the sample data using Gaussian Mixture Model along with Extreme Learning Machine (ELM). The system will emphasize on the speed of the response generated and it will be programmed to achieve higher accuracies to provide the users with a swift and quick human-like response generating system which they can put to effective use for leisure as well as emergency purposes.

II. RELATION TO PRIOR WORK

As applications of speech emotion recognition are increasing there are many projects aimed at recognizing speech emotion accurately. Various parameters have been used to train the system to detect emotion. Parameters like pitch, duration, intensity, formant, rhythm etc. are used in various combinations to provide quick and accurate results [1]. But more advanced features like Mel frequency cepstral coefficients (MFCC) proved to give more accurate results. And along with Gaussian Mixture models (GMM) the results are very close to expected outputs for a real life interactive system [2]. Thus to improve our results we are using a combination of MFCC and pitch based features. These features will be the input to GMM. GMM provides accuracy of nearly 76%, whereas hidden Markov model gives 71%, k-nearest neighbours gives 67%, feed-forward neural networks give 55% [3]. Thus by combining MFCC and pitch features we aim to provide comparatively higher accuracy.

III. ALGORITHM DETAILS

In this section, we describe our approach to classify emotions in speech. Speech in the training set is first segmented into frames of 20ms and from these frames the features are extracted. These features are fed into the expectation maximization algorithm to construct a Gaussian Mixture Model which results in various clusters. Testing data is then made to fit to one of the cluster based on its feature vector. The cluster indicates the emotion embedded in the speech and based on this an output is given. As our system is designed keeping in mind an in-car assistant, an output of either an appropriate location or a song depending on the atmosphere in the car is given. After some interval of time the mood in the car is again detected to analyse if the output has generated the desired result in the mood of the car. If the desired result is generated the priority of that output is increased, else it's priority is decreased.

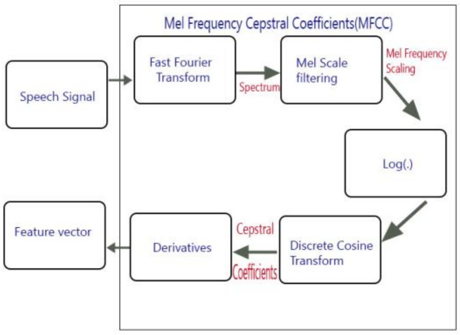

Figure 1

3.1 Feature Extraction

The speech signal is an analog signal, we need to process that signal to generate useful features that can be used to generate output. For the purpose of classifying emotions certain features in speech have been selected, the selection is based on an extensive study based on the performance of using different features to classify emotion. Pitch is one of the important features to determine emotion within speech. It is a highly gender dependent feature [4]. Pitch will be extracted using the VOICEBOX toolbox using a frame length of 20ms. If no voice is detected pitch is set to zero for that frame. Mel Frequency Cepstral Coefficients are widely used in speech recognition and speaker recognition applications. The MFCC features greatly represent human hearing. We need to take a sample of 20ms of sound and find its MFCC. First the power spectrum of the frame is calculated. The speech is then transmitted to Mel filterbanks that make it difficult to distinguish two frequencies especially when their values are high. This is also observed in humans. The logarithm of the filterbank energy is calculated to make the amplitude of sound non-linear. We finally calculate DCT of the log of filterbank energies which decorrelates the energies. Out of which only 12 coefficients of the DCT are kept as the higher ones will degrade performance. Further the mean, standard deviation of the feature and its derivative is calculated. Also the maximum, minimum and the range for each of the 12 features is calculated. The process is depicted in figure1. All the features are not used and only the features that give a better performance are selected [4].

Energy represents the intensity or volume of speech. It provides enough information to distinguish various emotions on its own but is not self-sufficient for all emotions. A frame size of about 20ms is used to sample energy.

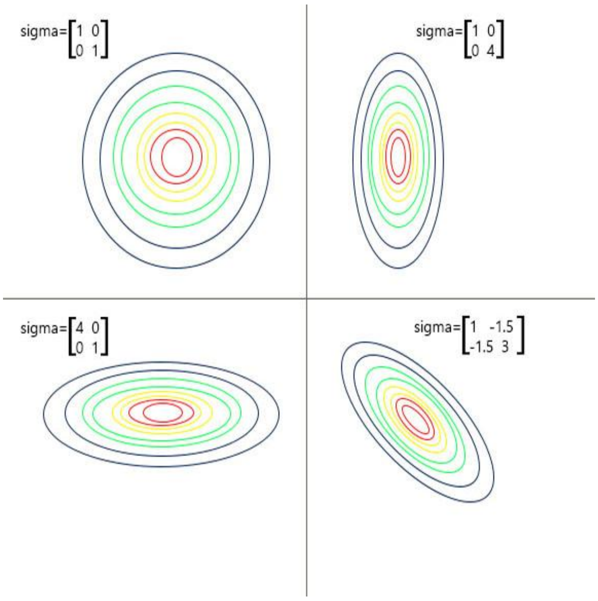

Figure 2

Hamming windows are used to sample the energy correctly. Short-time energy is given by,

$$E_n = \sum [\mathrm{x(m)}.\,\mathrm{w(n-m)}]^2,$$

Where w(n) is the window size.

3.2 The GMM Algorithm

The core of the project is based on the classification by Gaussian mixture models. The accuracy of our approach depends largely upon correct classification. GMM is a probability distribution which depends on the GMM component densities. Each vector will be given partial membership to a cluster by comparing the characteristics of the cluster and vector. The cluster which the vector has maximum membership in at the end of the algorithm is selected as the winning cluster and the emotion associated with it is given as output. Thus we see a distinct similarity between K-means clustering and GMM. The main difference between the two approaches is that K-means provides hard clustering. So a given vector belongs completely to a cluster as obtained by K-means algorithm. But in GMM a given vector may have partial membership in several clusters. Also, GMM clusters take into consideration the mean and variance of the components of the cluster while K-means depends solely on Euclidean distances from the centre of the cluster. Thus clusters of various shapes can be formed to classify input vectors as shown in figure 2.

To find out the cluster to which an input vector belongs, we use the Maximum Log likelihood function. The formula for computing the density for Gaussians contains exponential terms. We need to differentiate that equation and equate it to zero to find out the maximum distribution. As log is a monotone increasing function we can take log on both sides to nullify the exponentiation.

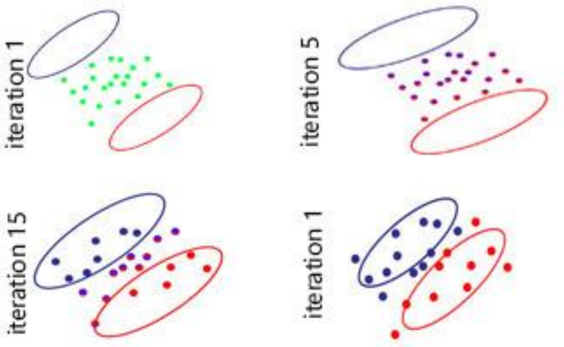

Figure 3

Before we begin the EM algorithm we need to initialize the clusters. The number of clusters and the number of emotions we need to classify should be the same. We can randomly initialize the clusters but EM is susceptible to local optima, thus we initialize them by using k-means in the first step and use the mean and variance of the clusters to set up the clusters for GMM. Once the clusters have been initialized we need to run EM iteratively on the training data, till convergence. In the expectation step, we attribute each input vector to the probability that it belongs to a given cluster as follows.

$$\mathrm{P(y{=}i}|x_j\,)\ \alpha\ \frac{1}{2\pi^{\frac{m}{2}}||\Sigma||^{\frac{1}{2}}}\ \mathrm{exp(A)\ P(y{=}i)}$$

Where,

$$\mathrm{A} = [-\frac{1}{2}(x_{j\,-}\mu_i)^T\,\Sigma_i^{-1}(x_j - \mu_i)^T]$$

The above is done for every cluster and each input vector. Thus we know the probability of a given sample belonging to a given cluster for this iteration. Summing the values P(y,x_j), we can calculate the total responsibility of a cluster. In the maximization step we use this responsibility to update the mean and covariance as follows,

$$m_c = \sum r_{ic} \ , \pi = \frac{m_c}{m}, \ \mu_c = \frac{1}{m} \sum r_{ic} x^{(i)}$$

$$\sum\nolimits_c = \frac{1}{m_c} \sum r_{ic} (x^{(i)} - \mu_c)^T (x^{(i)} - \mu_c)$$

Where r_{ic} is the responsibility of a cluster. Thus the mean and variance of the clusters get updated and the cluster gets shifted to represent its examples better in every iteration. Both the expectation and maximization steps are performed iteratively till they fit our data accurately.

3 Utterance Level Features

The above algorithm will give us the output for a 20ms frame. It is imprudent to detect emotion for such a short frame as the emotion will vary with time. Thus we now transform emotion recognition to a sequence classification problem. The emotion for each frame is a segment level emotion. We know need to find out the emotion for an entire utterance which we can assume is 30 seconds long. Thus we need to map for each 20ms frame to a 30 second utterance. For this we will assign use an Extreme Learning Machine (ELM) which is a modification of a Single Layer Feedforward Network. The ELM. An ELM has a single hidden layer and gives good performance [5] and is fast and has good classification characteristics [6].

An ELM generally uses more number of hidden nodes than in simple neural networks. The weights between the input layer and hidden layer remain constant and are assigned randomly. While the weights of the hidden and output layer are determined analytically. No iterative tuning is required for ELM hence they give a quicker response.

The algorithm for ELM is as follows.

1. Randomly assign weights a_i and b_i for hidden nodes.
2. Calculate hidden layer output matrix H= σ (W^T x_i), where is sigmoidal function.
3. And set output weight vector to $H^{\dagger}T$, where H is the Moore-Penrose generalised inverse of matrix H ($H^{\dagger}$ = $(H^T H)^{-1} H^T$) and T is target vector.

Thus, we observe that ELM uses time only to calculate the Moore-Penrose constant it is generally fast. The performance and theory of ELM is well explained in [6].

IV. EXPECTED RESULTS

In this paper we have proposed to use GMM algorithm along with extreme learning machine to detect the emotion of the conversation in the vehicle and generate responses accordingly. As per [3] the use of GMM boosts the performance of speech emotion recognition. Also [1] identifies the most prosodic features in speech. We combine the two studies to achieve higher learning accuracy. Further as stated in [5] extreme learning machine provides for faster and better performance. Thus, we propose to use extreme learning machine to classify the utterance level emotion whose probabilities of belonging to a particular class of emotion are initially calculated by GMM.

We expect that this combination can lead to an accuracy increment of approximately 20% as compared to the state of art algorithms when used individually. This in turn will increase the precision and utility of the response generated by the in-car system and give more appropriate output that becomes more efficient as the system is used as it learns to classify data better.

V. CONCLUSION

In this paper we have proposed to use GMM along with extreme learning machine to guarantee better results. GMM alone provides good classification capabilities and this hybrid model will definitely enhance the results. Thus this system is promising and opens doors for further research on the subject of enhancing mood of in-car passengers for better driving experience.

REFERENCES

[1] The production and recognition of emotions in speech: features and algorithms, Pierre-Yves Oudeyer (Sony CSL Paris, 6, rue Amyot, 75005 Paris, France)

[2] Emotion Recognition in Spontaneous Speech Using GMMs, Daniel Neiberg (1), Kjell Elenius(1) and Kornel Laskowski(2)((1) Department of Speech, Music and Hearing, KTH, Stockholm, Sweden (2) School of ComputerScience, Carnegie Mellon University, Pittsburgh, PA, USA)

[3] Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,,El Hayadi (Lab. of Pattern Anal. & Machine Intelligence, Waterloo Univ., Ont., Canada), Kamel, Karray.

[4] Classification on Speech Emotion Recognition- A Comparative Study, Theodoros Iliou, Christos-Nikolaos Anagnostopoulos (University of Aegean).

[5] Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine, Kun Han (Department of Computer Science and Engineering, The Ohio State University), Dong Yu, Ivan Tashev (Microsoft Research, One Microsoft Way)

[6] Extreme Learning Machine for Multilayer Perceptron Jieziong Tang, Student Member, IEEE, Chenwei Deng, Senior Member, IEEE and Guan-Bin Huang, Senior Member, IEEE.